CANADIAN CLIMATE CENTRE

REPORT NO. 89-7

REVIEW AND ASSESSMENT OF PROCEDURES FOR EXTREME VALUE ANALYSIS FOR SELECTED GEOPHYSICAL DATA

PHASE III
GUIDELINES AND RECOMMENDED METHODOLOGY FOR UNDERTAKING EXTREME VALUE ANALYSIS

ATMOSPHERIC ENVIRONMENT SERVICE

DOWNSVIEW, ONTARIO

1989

OTHER CANADIAN CLIMATE CENTRE (CCC)
SCIENTIFIC SERIES INCLUDE:

1. **CS series** - **CCC Climatological Study**
2. **CLI series** - **CCC Climatological Memoranda**
3. **CDS/DS** - **CCC Climatic Data Summary**
   - **CCC Documentation Sheet**

For further information on climatological publications contact the

Climatological Services Division
Atmospheric Environment Service
4905 Dufferin Street
Downsview, Ontario M3H 5T4

REVIEW AND ASSESSMENT OF PROCEDURES FOR EXTREME VALUE ANALYSIS FOR SELECTED GEOPHYSICAL DATA

PHASE III
GUIDELINES AND RECOMMENDED METHODOLOGY FOR UNDERTAKING EXTREME VALUE ANALYSIS

"This work was Supported by The Federal Panel on Energy R & D"

W.F. BAIRD & ASSOCIATES COASTAL ENGINEERS LTD.
HYDROTEK Water Resources Consultants Inc.
and
J.F. Lawless

March 1987

A publication of the Canadian Climate Program

DOWNSVIEW, ONTARIO

SUMMARY

This report is the third in a series of three reports describing investigations undertaken to develop a methodology for extreme value analysis for selected geophysical data. The first report was published as Canadian Climate Centre report No. 86-12 and consisted of a review of statistical procedures for estimating extreme values. The second report was published as Canadian Climate Centre Report No. 87-5 and described investigations undertaken to define the behavior and suitability of existing statistical models. Based on the results, specific recommendations for a general methodology were developed.

This report outlines the recommended methodology described in the Phase II report and summarizes the necessary equations and procedures. Worked examples are provided and additional references are provided for the user who requires more details or has a more complex problem.

This guideline has been written with the intended purpose of providing procedures which can be used by individuals undertaking estimates of extreme values without the benefit of computer based software and therefore does not incorporate all of the tests which might be undertaken in those circumstances. Readers who may have more sophisticated needs or who wish to develop computer based procedures are referred to the Phase I and II reports for guidance.

RÉSUMÉ

Il s'agit ici du troisième rapport d'une série de trois rapports qui décrivent les études entreprises pour concevoir une méthode d'analyse des valeurs extrêmes pour les données géophysiques sélectionées. Ce premier rapport, publié sous la désignation de rapport n° 86-12 du Centre climatologique canadien, consiste en un examen des méthodes statistiques d'estimation des valeurs extrêmes. Le deuxième rapport, publié sous la désignation de rapport n° 87-5 du Centre climatologique canadien, décrit les études entreprises pour définir le comportement et la pertinence des modeles statistiques actuels. Les résultats ont servi à concevoir des propositions particulières de méthodologie générale.

Le présent rapport fait état de la méthodologie proposée, décrite dans le rapport de la phase II, et résume les équations et les méthodes nécessaires. On fournit des exemples pragiques, ainsi que d'autres références pour l'usager qui a besoin de plus amples renseignments ou qui affronte un problème plus complexe.

Les présentes lignes de conduite visent à fournir des méthodes que peuvent suivre les personnes qui estiment les extrêmes sans disposer d'un logiviel informatique. Par conséquent, elle n'incorporent pas tous les essais qu'on pourrait effectuer en ces circonstances. On renvoie aux rapports de la phase I et de la phase II les lecteurs dont les besoins sont plus complexes ou qui désirent établir des méthodes informatiques.

TABLE OF CONTENTS

SUMMARY i
TABLE OF CONTENTS iii
ACKNOWLEDGEMENTS iv
1.0 INTRODUCTION 1
2.0 SELECTION OF EXTREME VALUE METHODOLOGY 3
2.1 Introduction 3
2.2 Case A, Record Length at Least 25 years 3
2.3 Case B, Record Length Between 10 and 25 years 3
2.4 Case C, Record Length Between 5 and 10 years 5
2.5 Case D, Record Length Less Than 5 years 5
3.0 RECOMMENDED METHODOLOGY FOR THE ANNUAL MAXIMUM (AMAX) APPROACH 6
3.1 Data Preparation and Pre-Analysis Examination 6
3.2 Estimation of Extremes 10
3.3 Post Analysis Checking 15
3.4 Uncertainty of Extremes Estimates 16
4.0 RECOMMENDED METHODOLOGY FOR THE PEAK OVER THRESHOLD (POT) MODEL 18
4.1 Data Preparation and Checking 18
4.2 Estimation of Extremes 22
4.3 Post Analysis Checking 25
4.4 Uncertainty of Extremes Estimates 27
5.0 ADDITIONAL CONSIDERATIONS 28
5.1 Choice of Underlying Variable 28
5.2 Data Uncertainty 28
5.3 Forensic Extreme Value Analysis 29
5.4 Other POT models 29
5.5 Multivariate Extreme Value Analysis 30
REFERENCES 31
APPENDIX A 33
APPENDIX B 37
APPENDIX C 43

ACKNOWLEDGEMENTS

This study was undertaken for the Atmospheric Environment Service under DSS Contract No. 01SE-KM147-4-3037.

The Scientific Authority for the study was Mr. Val Swail, Atmospheric Environment Service, CCAH, Downsview, Ontario and the work was reviewed by Mr. Les Welch, AES, Mr Bill Hogg, AES and Dr. Langley Muir, COGLA. Their interest and comments were appreciated.

Project manager for W.F. Baird and Associated Ltd. was Mr. J.S. Readshaw, P. Eng. Project Manager for HYDROTEK Inc. was Dr. W.E. Watt, P.Eng. and Mr. Ander Chow undertook many of the calculations required during the study. Dr. J.F. Lawless provided invaluable scientific and statistical guidance and his contribution is gratefully acknowledged.

This work was supported by the Federal Panel on Energy Research and Development.

1.0 INTRODUCTION

This report describes the third and final phase of an investigation undertaken to define a general methodology for extreme value analysis of selected geophysical parameters. During this three phase study, numerous investigations and tests of procedures were undertaken. In all cases, maritime wind data, recorded and archieved by AES were used. The procedures outlined in this report are applicable to most geophysical parameters.

In the first phase of the study an extensive review of existing procedures for univariate extreme value analysis was undertaken. It was found that three basic models existed for estimating extreme values:

- the time series model,
- the peak over threshold model, (POT),
- the annual maximum series models (AMAX).

The models differ primarily in the way the data series for the analysis are generated. The time series model is not widely used, primarily because of its complexity, but the remaining models are widely used by various disciplines. In the POT and AMAX models, there is a wide variation in specific procedures for data requirements, statistical distributions, methods of fitting, tests of fit and definition of uncertainty. In general, there is widespread agreement that any one of several procedures may be valid, but very little agreement regarding which one is best, or even appropriate, in any given situation. Further details of these procedures are provided in the Phase I report (WFBA-HYDROTEK, 1986) published as Canadian Climate Centre Report 86-12.

In the second phase of the study, selected long term data sets were used to investigate the behavior of the POT and AMAX models and to develop an understanding of which, or when, either model might be an appropriate choice for the estimation of extreme values. The results, which are reported in the Phase II report (WFBA-HYDROTEK, 1986), published as Canadian Climate Centre Report 87-5, indicated that:

- Data selection and screening were important elements of an extreme value analysis procedure and data should be reviewed for data quality, trend, jump, correlation and non-stationarity.
- In 5 out of 6 long term data series, the AMAX model, using an EV1 distribution, fit by the method of moments, provided an efficient and unbiased estimate if the number of years of record exceeded 25.
- The POT model generally provided unbiased estimates for record lengths of 15 years or more.
- When the number of years of record was less then 15, the POT model tended to yield more efficient estimates of extreme quantiles than the AMAX model.
- Regional information provided an invaluable adjunct to any analysis when less than 10 years of data were available.

Specific results for each model are summarized in the Phase II report (W.F. Baird & Associates - Hydrotek, 1987), published as Canadian Climate Centre Report 87-5.

In the third phase of the study, described in this report, general methodologies which provide sound, consistent and reliable estimates of extreme values for most geophysical data are outlined for both the AMAX and POT models, in Chapters 3 and 4 respectively. Guidelines for the choice of either model are provided in Chapter 2. In addition to the choice of model and the details of each methodology, there are a number of factors which should be considered prior to undertaking an extreme value study or should be reviewed during any such study. Factors which were identified during this study, but which are not fully discussed in the technical literature are discussed briefly in Chapter 5 as guidance for the interested user. All of these topics warrant further investigation.

The procedures described in this report are intended for individuals who do not have access to computer based software packages which may undertake all or parts of the analyses described in this report. Details of specific statistical tests are provided, together with worked examples, in the Appendices of this report. Additional references for tests are provided in this report.

Commercial and institutional software packages for general statistical and specific extreme value analysis are available. The user should review both the default and optional procedures in these programs, prior to their use, to determine if they conform with the recommendations of this report.

2.0 SELECTION OF EXTREME VALUE METHODOLOGY

2.1 Introduction

The results of the Phase II study indicated that the choice of appropriate model for undertaking extreme values analysis depended primarily on the number of years of data which were available. Once a model (AMAX or POT) is selected on this basis, the details of the appropriate methodology can be defined, based on the experience in Phase II. These procedures were developed specifically for winds, but based on underlying theoretical considerations they are also applicable for other data types when long-term data sets are not available.

An outline of a recommended approach for estimating extreme values is shown in Figure 2.1. The relevant chapters which describe the detailed methodology are also indicated and the user can refer to the appropriate section for guidance. Worked examples are provided in Appendices and a detailed discussion of the basis behind each procedure is provided in the Phase I report.

2.2 Case A, Record Length at Least 25 years

When at least 25 years of record are available, experience has indicated that the Annual Maximum (AMAX) model provides sufficiently reliable estimates of extreme values, provided that:

- appropriate data preparation and checking procedures are successfully completed,
- appropriate distributions are selected and their parameters are estimated correctly

The recommended methodology for the AMAX model is described in Chapter 3.0

2.3 Case B, Record Length Between 10 and 25 years

When between 10 and 25 years of record are available, experience has indicated that either the AMAX or the POT models may provide sufficiently reliable estimates of extreme values provided that:

- appropriate data preparation and checking procedures are successfully completed,
- appropriate distributions are selected and their parameters are estimated correctly

The recommended methodology for the AMAX model is described in Chapter 3.0.

The recommended methodology for the POT model is described in Chapter 4.0.

In general the POT model gives better results when significantly less than 25 years of data are available. The POT model estimates tend to be unbiased provided that the number of

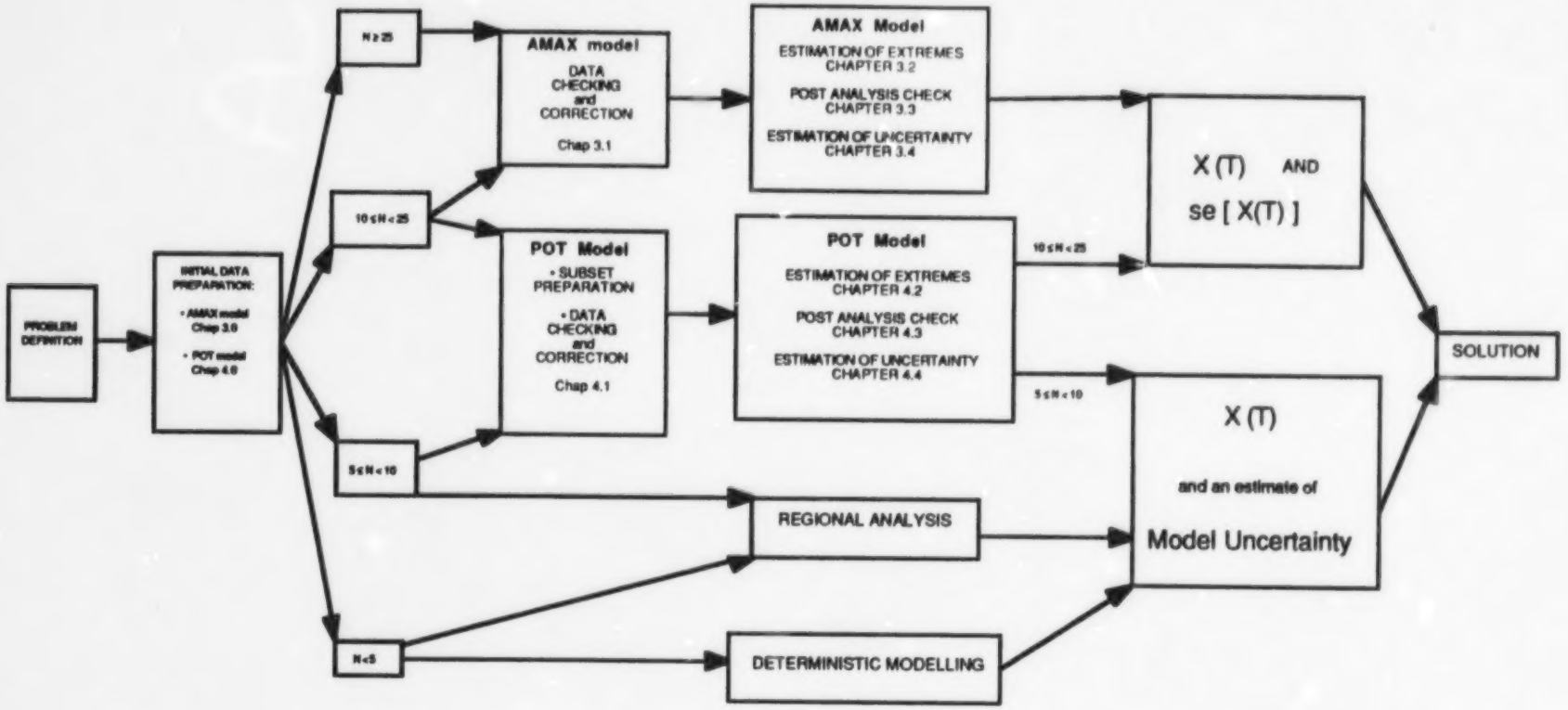

FIGURE 2.1

OUTLINE OF RECOMMENDED METHODOLOGY

years of record are greater than 15. When the number of years (N) of record are less than 15, the estimates may be only slightly biased high or low if at least 2N events are selected. As the number of years of record approaches 25 both the AMAX and the POT models give unbiased estimates, but the standard error of estimate (se_x) is smaller for the POT model.

As a guideline when two estimates of an extreme event are produced, the results of the AMAX model should be consistent within acceptable limits defined by the POT standard error, before it is accepted as a reliable estimate.

2.4 Case C, Record Length Between 5 and 10 years

When between 5 and 10 years of record are available, experience has indicated that the POT model provides sufficiently reliable estimates of extreme values, provided that:

- appropriate data preparation and checking procedures are successfully completed,
- appropriate distributions are selected and their parameters are estimated correctly
- the number of events selected is at least 20.

The recommended methodology for the POT model is described in Chapter 4.0.

A regional analysis can also provide useful support for the estimates of extremes developed using the POT model if the data are available. The regional analysis should be undertaken following the appropriate procedures described in this report. Additional material describing procedures for a regional analysis for estimates of discharge, also in Hogg and Carr (1985) which discusses regional analysis for rainfall and in the National Building Code which discusses regional analysis for winds.

2.5 Case D, Record Length Less Than 5 years

When less than 5 years of data are available, a combination of regional analysis and deterministic modelling should be used to develop estimates of extreme values. The definition of appropriate procedures for undertaking either type of investigation was beyond the scope of this report and should be undertaken with the assistance of appropriate expertise.

Based on the studies described in the Phase II report, there is some evidence to support the conclusion that the ratio of the 100 year return period windspeed, and the mean windspeed, is approximately 1.5 near Canadian Marine waters.

3.0 RECOMMENDED METHODOLOGY FOR THE ANNUAL MAXIMUM (AMAX) APPROACH

This chapter describes the recommended procedure when the AMAX model is the appropriate extreme value model as described in Chapter 2.0. A worked example following these guidelines is provided in appendix A.

3.1 Data Preparation and Pre-Analysis Examination

3.1.1 Data Preparation

The following steps should be undertaken for the preparation of data for an extreme value analysis:

- Obtain the initial data directly from the appropriate data collection agency. It is assumed that annual maxima data are available. If this is not the case, and the data must be developed from a more frequent time series (such as hourly windspeed or all incidents of icing), the procedures outlined in Chapter 4.1.1 should be adopted.

- Obtain the data station description and history. Details which should be requested and included with a station report include:

  - A description of the geographical setting of the station including instrument location
  - Instrument type and installation details, (height above ground, clearance from nearby obstacles etc)
  - Details of calibration history and maintenance cycle
  - Details of any post-processing and data quality checks which might have been undertaken by the data collection or archiving agency.

- Check the data record for completeness. A data series is often incomplete, usually for one of four reasons:

  - Data collection may have been discontinued for a period because of financial or manpower constraints,
  - Data may be missing because recording was affected by the phenomena of interest
  - Data may be missing because of the archiving procedure used by the data collection agency.
  - Data may appear to be missing because of the presence of zeros

In the first case, the recommended solution is to concatenate the record as illustrated in Figure 3.1, so that the period of record becomes equal to the total duration of available data. Care should be exercised where the missing data might include a period of significant activity.

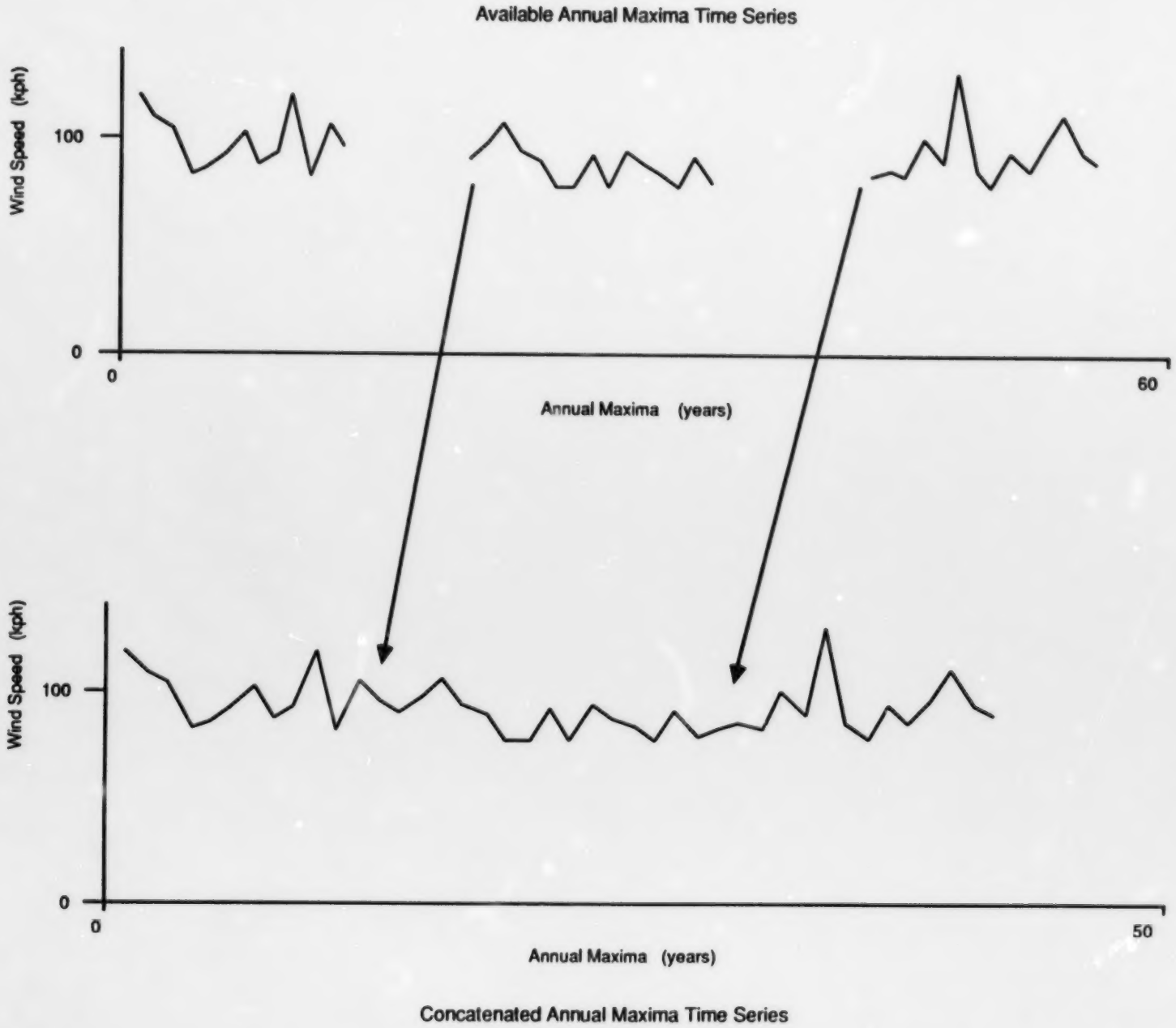
Available Annual Maxima Time Series
Wind Speed (kph)
100
0
0
60
Annual Maxima (years)
Wind Speed (kph)
100
0
0
50
Annual Maxima (years)
Concatenated Annual Maxima Time Series

FIGURE 3.1

In the second case, for example, where the windspeed was greater than the capability of the instrument to record it, the responsible data agency should be requested to assist in estimating the missing data to complete the record.

In the third case, it may be necessary to obtain additional information on the procedures followed during archiving.

In the fourth case, most data agencies differentiate between zeros and missing data by coding the different reasons. If this is not the case, then it must be determined whether differences exist and whether the data is missing. The record can then be concatenated as described in Figure 3.1. If the missing data is due to a true zero, only the conditional probability can be determined, given that the process of interest occurs.

3.1.2 Pre-Analysis Examination

As the first step in the pre-analysis examination, the annual maximum series should be plotted as a time series as shown in Figure 3.2. This type of plot may indicate that trend, jump, lack of independence or homogeneity is present in the data set. The existence of any of these characteristics should be verified using the statistical tests outlined in Table 3.1.

Trend would be indicated by a tendency for the moments (mean, covariance, etc.) of the sample to change gradually but distinctively over time.

Jump would be indicated by a tendency for the moments to make an abrupt change over time.

Lack of independence would be indicated, for example, by a trend for a high to follow high, or a low to follow low, or a high to follow very low or vice versa.

Non-Homogeneity may be difficult to determine as an understanding of the processes is required. A suggested approach to examine for non-homogeneity would be to examine details of the five highest and the five lowest events in a series in order to determine if the underlying processes responsible for the observations are different from the remainder of the sample observations.

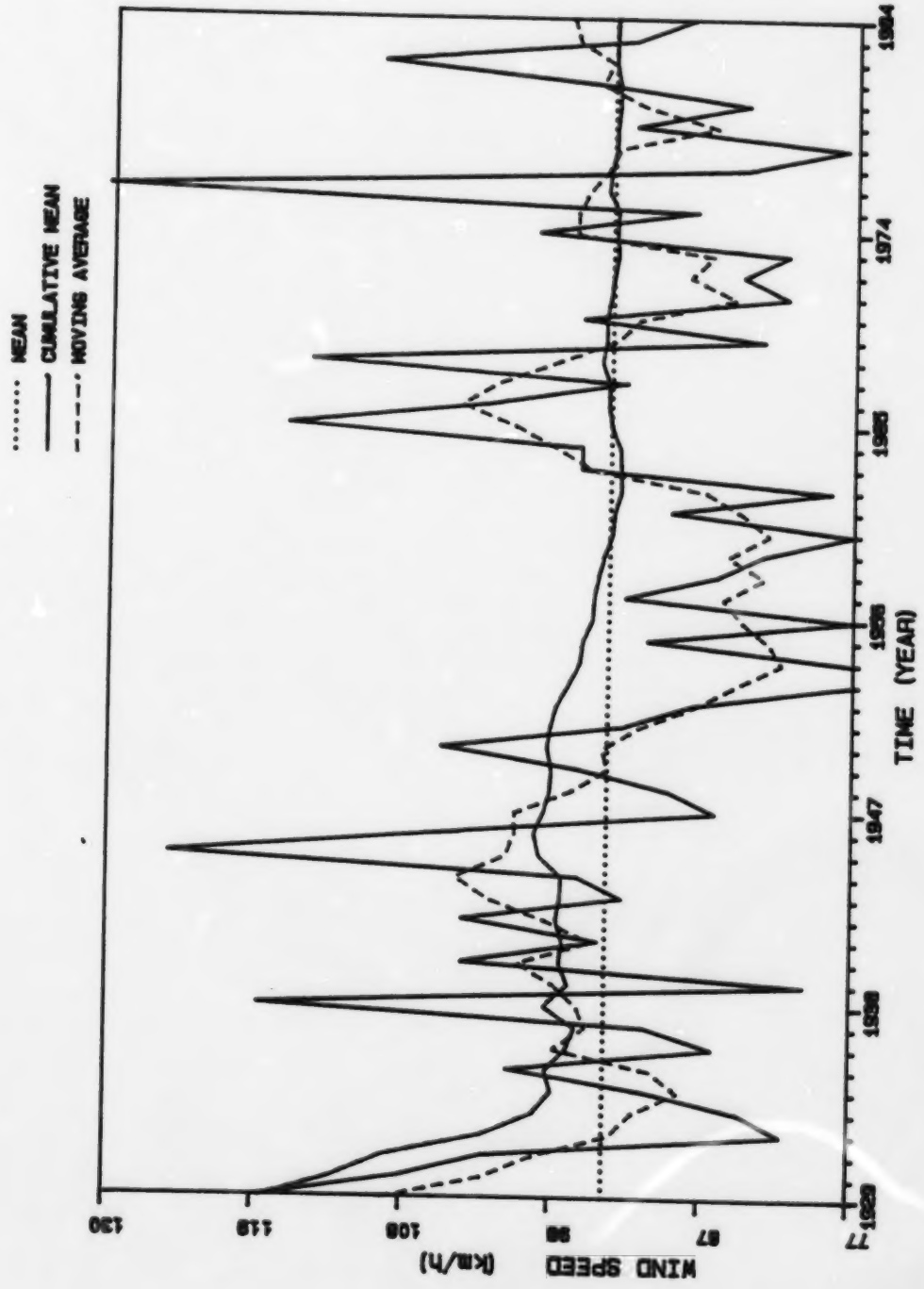
MEAN
CUMULATIVE MEAN
MOVING AVERAGE
WIND SPEED (KM/H)
77
87
98
108
118
130
1929
1938
1947
1956
1965
1974
1984
TIME (YEAR)

FIGURE 3.2

Table 3.1
Statistical Tests for Pre-Analysis Examination

Characteristic	Data Set	Test	Additional Reference
Trend	Complete time series correlation coefficient	• Spearman rank order	Appendix C 1
Jump	Two time series: • Before jump • After Jump	• Wald-Wolfowitz • Mann Whitney test for jump	Appendix C2 Appendix C3
Independence	Complete time series	• Auto correlation function • Spearman rank order serial correlation coeff.	Appendix C4 Appendix C5
Homogeneity	Two sub-series: • Process A • Process B	• Wald-Wolfowitz • Mann Whitney for homogeneity	Appendix C6 Appendix C7

Where these tests provide a strong indication of trend, jump, lack of independence or non-homogeneity, further analysis is required. The appropriate procedures will vary in each case and should be undertaken with the assistance of appropriate expertise.

3.2 Estimation of Extremes

3.2.1 Calculation of Sample Moments

From a sample of size N corresponding to N observations of annual maxima x_i, compute the sample mean $\bar{x}$, the sample standard deviation s_x, and the sample coefficient of skewness g_x.

$$\bar{x} = \left(\sum_{i=x}^{N} x_i \right) / N$$

$$s_x = \left\{ \sum_{i=x}^{N} \left(x_i - \bar{x} \right)^2 / (N-1) \right\}^{\frac{1}{2}}$$

$$g_x = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \left(x_i - \bar{x} \right)^3 / \left(s_x \right)^3$$

3.2.2 Graphical Presentation of Observed Annual Maxima

a) Arrange the N annual maxima x_i in order of decreasing magnitude and assign a rank m (m = 1 for largest value).

b) For each x_i compute a plotting position, p_m, using the Cunnane formula:

$$p_m = (m - 0.4) / (N + 0.2)$$

and the corresponding "empirical return period":

$$T_m = 1/p_m$$

c) Plot each x_i at its corresponding plotting position on Extreme Value Type I (EV1 or "Gumbel") probability paper. Points should be relatively large, typically of the order of 1/80 of the minimum plot dimension. The legend should note "observed annual maxima". An example of EV1 probability paper is shown in Appendix A7 of this report. If probability paper is not available the data can be plotted on linear graph paper using the linear reduced variate, Rv, for plotting location, where:

$$Rv = - \ln (-\ln (1 - p_m))$$

3.2.2 Selection of Distribution

a) For $N \leq 25$, assume an EV1 distribution and proceed to section 3.2.4.

b) For $N > 25$, compare the skewness coefficient g_x with the appropriate confidence limits for an EV1 distribution. (Shown in Figure 3.3).

The lower limit of skewness = $g_l = 0.5 + 0.15 (\sqrt{N} - 5)$ for $N > 25$.

The upper limit of skewness = $g_u = 1.1 + 0.15 (\sqrt{N} - 5)$ for $N > 25$.

If:

$$g_l \leq g_x \leq g_u,$$

assume an EV1 distribution and proceed to section 3.2.4.

If:

$$g_x < g_l \text{ or } g_x > g_u,$$

there are two likely explanations for the high or low skewness

- it may result from one or more high or low outliers which can be identified in the post-analysis examination described in section 3.3; or,
- it may be because the parent distribution is a generalized extreme value distribution, GEV, and not EV1.

In the second case, assume a GEV distribution and proceed to section 3.2.5.

Sample Skewness Coefficent for the EV1 Distribution

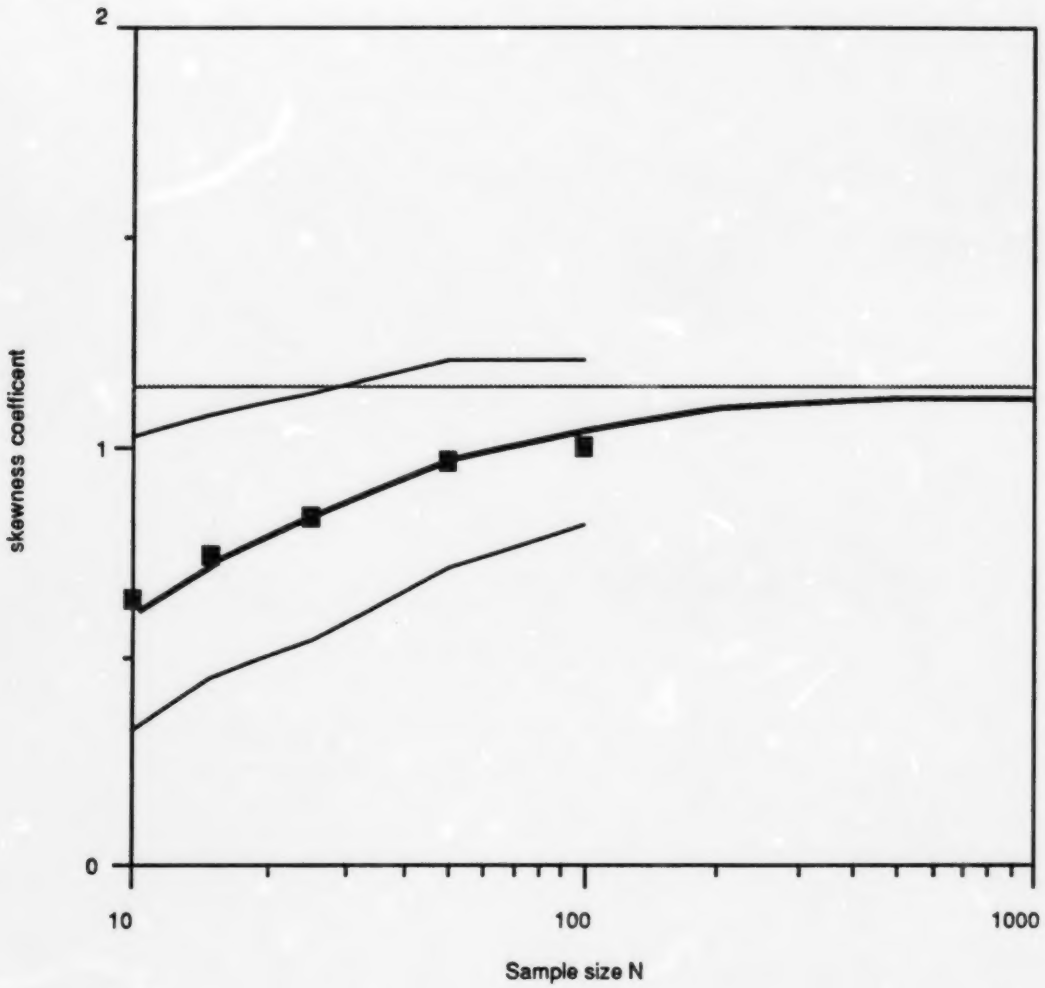

FIGURE 3.3

3.2.4 Quantile estimates for EV1 Distribution

Parameters of the EV1 distribution are estimated by moments so there is no need to compute the two parameters explicitly. Instead, the parameters are implicitly defined by ($\bar{x}$) and s_x, and the frequency factor K_T. Quantiles are estimated using:

$$x_T = \bar{x} + K_T s_x$$

where: $$K_T = -(\sqrt{6}) [0.577 + \ln (\ln (T/(T - 1)))] / \pi$$

Values of K_T for common periods are tabulated in Table 3.2.

Table 3.2
Frequency Factors for EV1 Distribution

Return Period, T	2	5	10	20	25	50	100
Frequency Factor, K_T	-0.164	0.719	1.304	1.866	2.044	2.592	3.137

3.2.5 Quantile Estimates for GEV Distribution

Parameters for the GEV distribution are estimated by moments, but there is no need to compute the three parameters explicitly. Instead the parameters are implicitly defined by $\bar{x}$ and s_x and the frequency factor $K_{T, g}$ which in this case depends on the coefficient of skewness. Quantiles are estimated using:

$$x_T = \bar{x} + (K_{T, g}) s_x$$

The formula for the frequency factor of the GEV distribution is complex and for simplicity, values of $K_{T, g}$ are given in Table 3.3 for commonly used return periods.

Table 3.3
Frequency Factors for GEV Distribution

Coefficient of Skewness g_x	kf *	Frequency Factor K(T,g) Return Period (T)						
		2	5	10	20	25	50	100
.00	.28	-.008	.870	1.316	1.665	1.762	2.025	2.240
.10	.25	-.026	.860	1.325	1.697	1.802	2.090	2.331
.20	.22	-.043	.850	1.331	1.726	1.838	2.152	2.420
.30	.19	-.060	.838	1.336	1.751	1.872	2.212	2.507
.40	.16	-.075	.826	1.338	1.773	1.901	2.267	2.590
.50	.14	-.089	.814	1.338	1.792	1.927	2.317	2.668
.60	.12	-.101	.802	1.337	1.808	1.950	2.362	2.739
.70	.10	-.112	.791	1.335	1.821	1.968	2.401	2.802
.80	.08	-.120	.781	1.333	1.831	1.983	2.433	2.855
.90	.07	-.127	.773	1.330	1.838	1.994	2.459	2.898
1.00	.06	-.132	.766	1.328	1.843	2.002	2.478	2.930
1.10	.00	-.164	.719	1.305	1.866	2.044	2.592	3.137
1.20	.00	-.164	.719	1.305	1.866	2.044	2.592	3.137
1.30	-.03	-.176	.695	1.286	1.864	2.049	2.627	3.211
1.40	-.04	-.183	.684	1.280	1.868	2.059	2.656	3.266
1.50	-.05	-.190	.671	1.271	1.869	2.063	2.678	3.311
1.60	-.07	-.195	.659	1.261	1.867	2.066	2.696	3.351
1.70	-.08	-.201	.648	1.252	1.865	2.067	2.712	3.388
1.80	-.09	-.205	.636	1.242	1.862	2.067	2.725	3.421
2.00	-.11	-.213	.615	1.223	1.854	2.064	2.747	3.478
2.10	-.12	-.216	.606	1.213	1.849	2.062	2.755	3.503
2.20	-.13	-.219	.596	1.204	1.844	2.059	2.762	3.525
2.30	-.13	-.222	.588	1.195	1.838	2.056	2.768	3.545
2.40	-.14	-.225	.579	1.187	1.833	2.052	2.773	3.563
2.50	-.15	-.227	.571	1.178	1.827	2.048	2.777	3.580
2.60	-.15	-.229	.563	1.170	1.822	2.044	2.780	3.595
2.70	-.16	-.231	.556	1.162	1.816	2.039	2.782	3.608
2.80	-.17	-.233	.549	1.154	1.810	2.035	2.784	3.620
2.90	-.17	-.234	.543	1.147	1.804	2.030	2.785	3.631
3.00	-.18	-.235	.536	1.140	1.799	2.026	2.786	3.640

* value of the parameter kf implied by g_x

3.2.6 Graphical Presentation of Quantile Estimates

Plot the quantile estimates, for range of return periods, on the same EV1 paper as the observed annual maxima are plotted. Join the points with a bold line labelled "fit line - EV1 (or GEV) distribution". Note that for the EV1 distribution, the fit line will appear as a straight line on EV1 probability paper, whereas for the GEV distribution the fit line will be a curve.

3.3 Post Analysis Checking

In many cases the existence of an apparent outlier may be suspected after the extreme value analysis is completed. An outlier may be suspected because one or more data points depart significantly from the trend of either the data or the fitted distribution plotted as described above. An outlier may exist for one of several reasons:

- it may deviate from other observations because it was generated by some other process
- it may be a rare event from the same population distribution
- it may be an observational error which was not detected in pre-analysis examination
- the statistical model may be wrong.

The extent to which an outlier may represent a problem is a function of the type of distribution, the sample size and the presence of corroborative regional information. It should be noted that:

- three parameter distributions, such as the GEV, are more sensitive to outliers than two parameter distributions such as the EV1.
- the influence of an outlier on the moments of a distribution are smaller for a large sample than for a small sample
- if the date of the outlier is known, raw or other data records can be reviewed for corroborative evidence which might support or explain the presence of the outlier.

3.3.1 High Outliers

If the apparent outlier is a high outlier, that is, it has a low probability of occurrence associated with it, the following procedures should be undertaken:

- Determine if the data point is correct, by checking the source or basis of the data for errors associated with either measurement, recording or archiving or methodology, especially where the data was derived using numerical models.
- Verify the physical processes which are the underlying cause of the observation and determine if it is consistent with the other observations.

If the outlier is the result of a non-homogeneous process it should be removed and consideration should be given to quantifying and combining the probabilities of two processes. If the apparent outlier is a homogeneous observation, the calculations should be rechecked and if they are correct:

- the observation may be a rare event, or,
- the model may be wrong, and consideration should be given to using another extreme value model.

If the model is not wrong or there is no other suitable extreme value model, then the best that one can do is to repeat the analysis without the outlier and define an envelope of likely return periods for the observation.

In the special case of Forensic analysis, discussed further in Chapter 5.3, a more detailed understanding of outlier and their statistical significance is required.

3.3.2 Low Outliers

The occurrence of low outliers is an often overlooked phenomenon, but the user should review any analysis for the presence of such outliers. In particular low outliers can significantly distort sample skewness and the user will be inclined to use distributions which are consistent with a low skewness (i.e. the Normal distribution) or a three parameter distribution such as the GEV, which can have an upper limit. The danger of a low outlier is that it may create an artificial case where the estimated upper limit is less than the largest observation.

If a low outlier is suspected, the following procedures should be undertaken:

- Determine if the data point is correct, by checking the source or basis of the data, for errors associated with either measurement, recording, archiving or methodology, especially where the data was derived using numerical models.
- Verify the physical processes which are the underlying cause of the observation and determine if it is consistent with the other observations
- If the outlier is the result of a non-homogeneous process it should be removed and consideration should be given to how the probabilities of two processes should be quantified and combined.
- If the apparent outlier is consistent with the remainder of the sample, the calculations should be rechecked and if they are correct, the model may be wrong. Consideration should then be given to using another extreme value model.
- If the GEV distribution is selected, then the estimated upper band should be checked to ensure that it is consistent with the observed data.

3.4 Uncertainty of Extremes Estimates

The uncertainty of the extremes estimates, due to sample size alone, can be approximated using the sampling standard error, $se(x_T)$. This can be combined with the normal distribution to estimate confidence limits about any estimate of a quantile, x_T.

3.4.1 Sampling Standard Error - EV1 Distribution

When parameters are estimated by moments, the sampling standard error $se(x_T)$ can be computed from

$$se(x_T) = (s_x / \sqrt{N})\, \delta_T$$

where $$\delta_T = (1 + 1.14\, K_T + 1.10\, K_T^{\,2})^{1/2}$$

Values of δ_T for common return periods are provided in Table 3.4.

Table 3.4
Standard Error Factors for EV1 Distribution

T (Years)	2	5	10	20	25	50	100
δ_T	0.918	1.546	1.951	2.638	2.815	3.368	4.025

3.4.2 Sampling Standard Error - GEV Distribution

When parameters are estimated by moments, the sampling error $se(x_T)$ can be computed from

$$se(x_T) = (s_x / \sqrt{N})\, \delta_{T,g}$$

There is no simple formula for $\delta_{T,g}$ and tables do not exist. As an approximation, the variation with skewness could be ignored and values given in Table 3.4 for the EV1 distribution (i.e. $g_x = 1.14$) could be used. The actual standard error will exceed this value because of the sampling error for the coefficient of skewness (g_x).

4.0 RECOMMENDED METHODOLOGY FOR THE PEAK OVER THRESHOLD (POT) MODEL

4.1 Data Preparation and Checking

The following steps should be undertaken for the preparation of data for a POT extreme value analysis:

- Obtain the initial data directly from the appropriate data collection agency.
- Obtain the data station description and history. Details which should be requested and included with a station report include:

  - A description of the geographic setting of the station including instrument location
  - Instrument type and installation details, (height above ground, clearance from nearby obstacles, etc)
  - Details of calibration history and maintenance cycle
  - Details of any post-processing and data quality checks which might have been undertaken by the data collection or archiving agency.

- Check the data record for completeness. A data series is often incomplete, usually for one of four reasons:

  - Data collection may have been discontinued for a period because of financial or manpower constraints,
  - Data may be missing because recording was affected by the phenomena of interest
  - Data may be missing because of the archiving procedure used by the data collection agency.
  - Data may appear to be missing because of the presence of zeros

In the first case the recommended solution is to repair and then concatenate the record as illustrated in Figure 4.1, so that the period of record becomes equal to the total duration of available data. Care should be exercised where the missing data period might reasonably be expected to include a period of significant activity.

Repair of the data may be undertaken with the assistance of appropriate expertise where a knowledge of the applicable process suggests that events of interest were not missed during the repaired gaps. Repair ensures that the record length is as long as can legitimately be inferred for the data set.

In the second case, for example, where the magnitude of the process of interest was greater than the capability of instruments to record it, the responsible data agencies should be requested to assist in estimating the missing data.

In the third case, data may be missing because the data were recorded in one set of units and either converted to another set or rounded. Due to metrification in the 1970s, it is possible for conversion and rounding to have occurred twice during the archiving process and as a result the occurrence of certain data in the archived record may be artificially impossible.

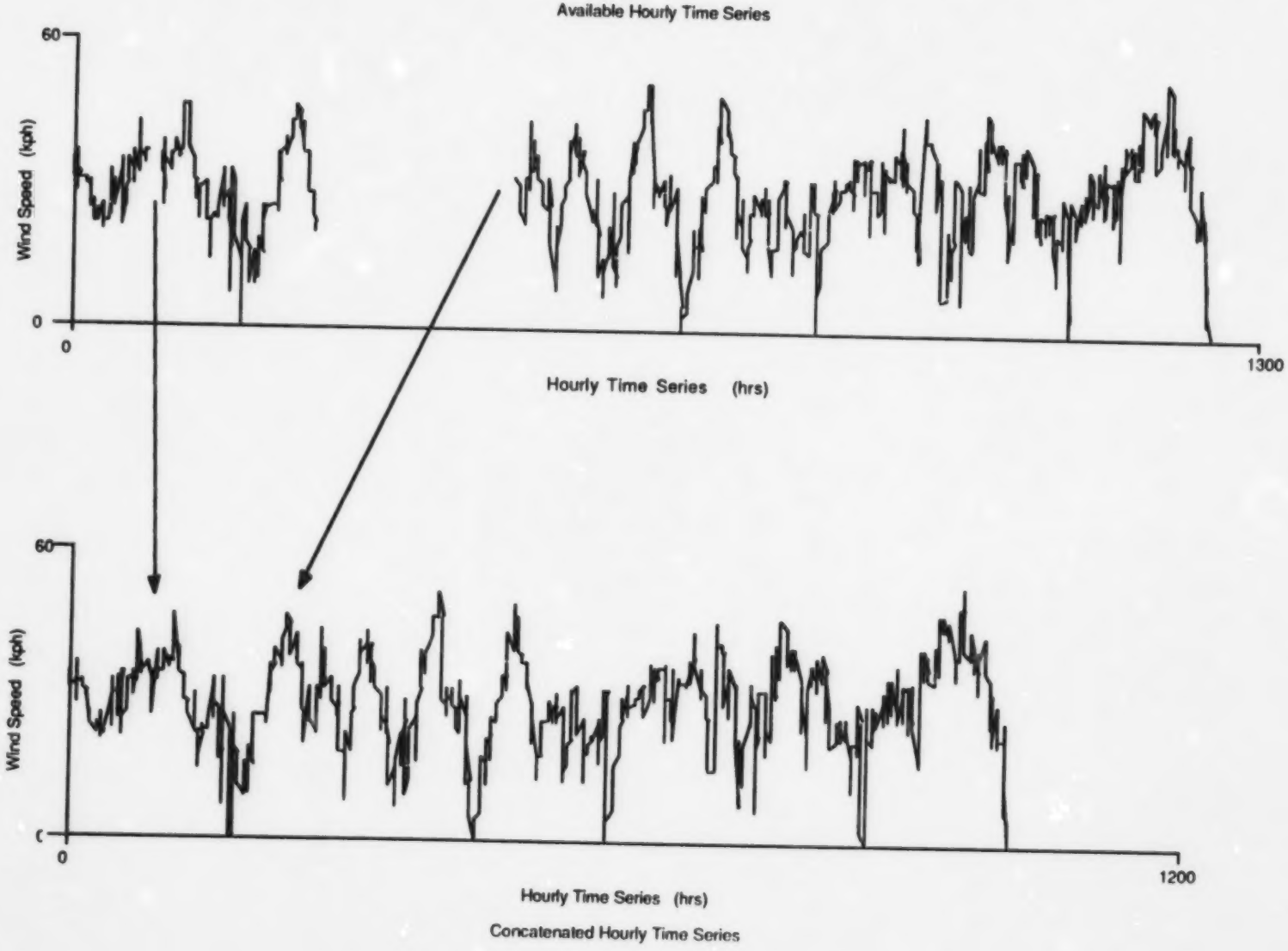
Available Hourly Time Series
60
Wind Speed (kph)
0
0
1300
Hourly Time Series (hrs)
60
Wind Speed (kph)
0
1200
Hourly Time Series (hrs)
Concatenated Hourly Time Series

FIGURE 4.1

In the fourth case, most data agencies differentiate between zeros and missing data by coding the different reasons. If this is not the case, the reasons must be determined and if the data are missing, then the record can be concatenated as indicated in Figure 4.1. If the missing data are true zeros, only the conditional probability, given that the process of interest occurs, can be determined.

- The data should be corrected as necessary as described above, and where sufficient information exists, it should also be corrected for any trend or jump caused by changes documented in the station history.

- The parent POT series should be selected in chronological order to satisfy the criteria defined below:

  - the threshold must be above the mode of the parent distribution
  - at least 20 events must be selected. It is advisable, provided that the threshold is greater than the mode, to select between 5 and 10 times N to define the initial parent POT series
  - there must be at least two events per year

- The data should be screened to ensure that a minimum inter-event time is exceeded. The minimum inter-event is required to ensure that selected events are independent of each other in a manner which is consistent with the knowledge of the geophysical process of interest. Selection of an inter-event time should be undertaken with the assistance of appropriate expertise.

- Arrange the resulting observations in order of decreasing magnitude and assign a rank m (m = 1 for the largest value).

4.1.2 Pre-Analysis Examination

As the first step in the pre-analysis examination, the selected POT series should be plotted as a time series as shown in Figure 4.2. This type of plot may indicate that trend, jump or lack of independence or homogeneity is present in the data set. The existence of any of these characteristics should be verified using the statistical tests outlined in Table 4.1.

Where these tests provide a strong indication of trend, jump, lack of independence or non-homogeneity, further analysis is required. The appropriate procedures will vary in each case and should be undertaken with the assistance of appropriate expertise.

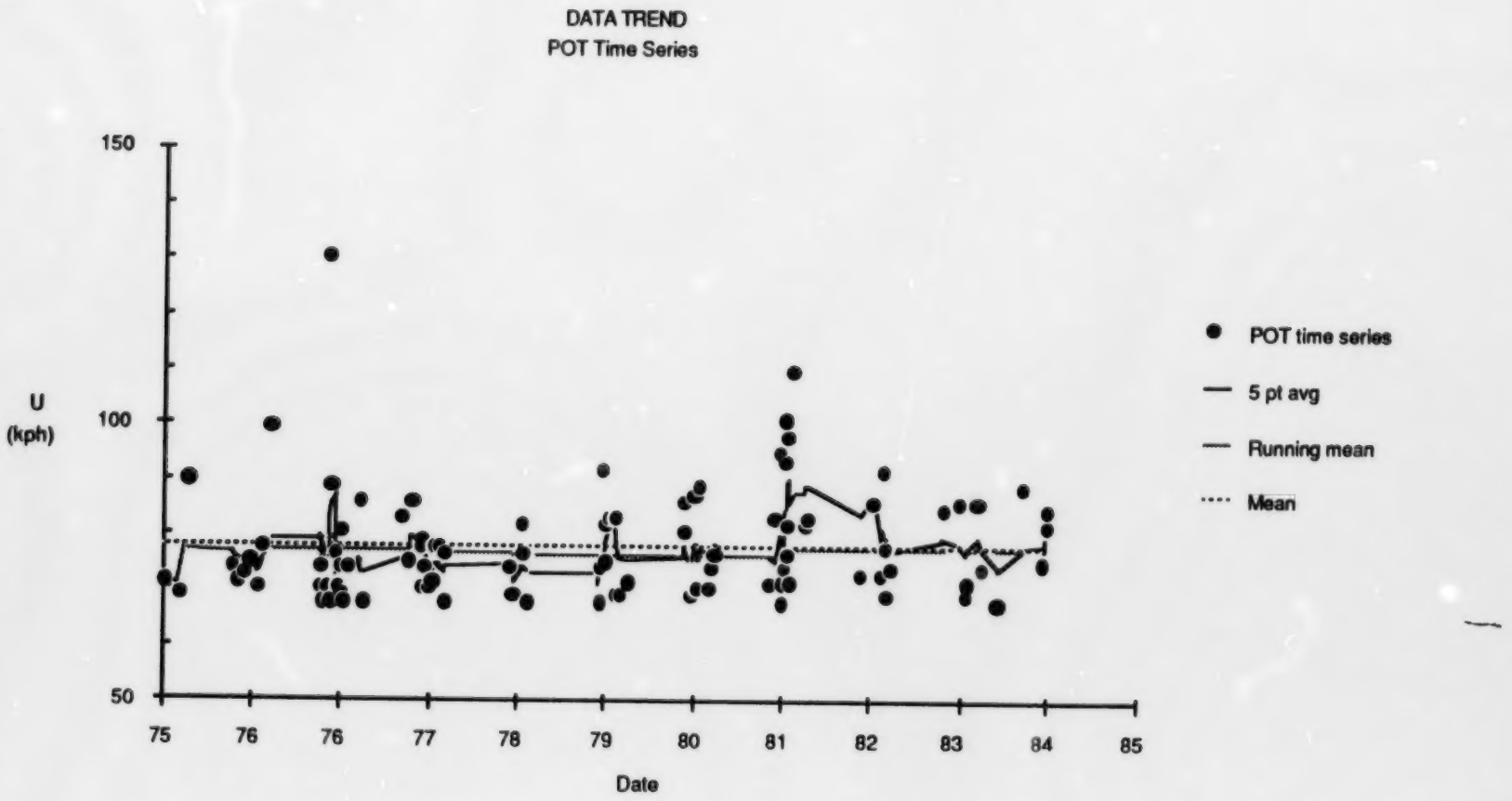
DATA TREND
POT Time Series
150
100
50
U
(kph)
75
76
76
77
78
79
80
81
82
83
84
85
Date
POT time series
5 pt avg
Running mean
Mean

FIGURE 4.2

Table 4.1
Statistical Tests for Pre-Analysis Examination

Characteristic	Data Set	Test	Additional Reference
Trend	Complete time series	• Spearman rank order correlation coefficient	Appendix C 1
Jump	Two time series: • Before jump • After Jump	• Wald-Wolfowitz • Mann Whitney test for jump	Appendix C2 Appendix C3
Independence	Complete time series	• Auto correlation function • Spearman rank order serial correlation coeff.	Appendix C4 Appendix C5
Homogeneity	Two sub-series: • Process A • Process B	• Wald-Wolfowitz • Mann Whitney for Homogeneity	Appendix C6 Appendix C7

4.2 Estimation of Extremes

4.2.1 Calculation of Sample Moments

From a sample of size M, corresponding to M observations with x_i (i = 1 to M) above a threshold x_0, compute the sample mean $\bar{x}$, the sample standard deviation s_x, and the sample coefficient of skewness g_x.

$$\bar{x} = \left(\sum_{i=x}^{M} x_i \right) / M$$

$$s_x = \left\{ \sum_{i=x}^{M} \left(x_i - \bar{x} \right)^2 / (M-1) \right\}^{\frac{1}{2}}$$

$$g_x = \frac{M}{(M-1)(M-2)} \sum_{i=1}^{M} \left(x_i - \bar{x} \right)^3 / \left(s_x \right)^3$$

4.2.2 Definition of the POT Parameters

Because the threshold is selected arbitrarily, and its choice affects both the resulting sample size (which can be large), and the values of the POT parameters, the relationship between the threshold and the resulting parameters must be checked in an iterative manner to ensure that the assumptions of the POT model are satisfied (i.e. a Poisson process and an exponential distribution of peaks).

The recommended procedure, therefore, is to select a threshold, estimate the parameters, check the POT model and repeat until either the POT model assumptions are satisfied, or the model is shown to be not suitable for the application at hand. After this point, the data can be checked graphically again for trend, jump, independence and homogeneity. The elements of this procedure are described below:

- The threshold x_0 is defined in the selection of the POT data series as described in section 4.1.1

- The Poisson distribution parameter λ is estimated by moments as

$$\lambda = M/N$$

  where M is the number of events and N is the record length in years.

- The two exponential distribution parameters are the threshold x_0 and the scale parameter β. The scale parameter is estimated by the method of moments as

$$\beta = \bar{x} - x_0$$

  Where the distribution of peaks is distorted because of gaps (i.e. by recorder bias) and where one of these gaps corresponds to the threshold, β should be estimated from the second moment using

$$\beta = s_x$$

There are two main assumptions underlying the POT model described above:

- For a Poisson process, the ratio of the mean to the variance of the number of events per year is unity.

- For an exponential distribution of magnitudes, the population coefficient of skewness is two.

Experience has shown that the results of a POT model are not sensitive to the Poisson process and therefore the second requirement is more important than the first. The validity of the POT model can be assessed by comparing the sample values to the two criteria. If they differ significantly, the threshold should be increased, and moments and parameters recalculated, until both criteria are satisfied and at least the sample skewness approaches two or the sample size M reaches a minimum allowable value of 2N or 20.

An additional useful test for the Poisson process is that the inter-event durations should be exponetially distributed.

4.2.3 Quantile Estimates

Quantiles are estimated using the following expression:

$$x_T = x_o + \beta \ln\lambda + \beta y_T$$

where the reduced variate $y_T = -\ln \ln (T/(T-1))$, values of y_T are given in Table 4.2 for common return periods.

Table 4.2

Reduced Variate for POT Quantiles

Return Period, T	2	5	10	20	25	50	100
Reduced Variate, y_T	–0.367	1.500	2.250	2.970	3.199	3.902	4.600

4.2.4 Graphical Presentation

a) For each x_i, compute a plotting position using the Gringorten formula (as an approximation of the exponential probability distribution plotting position):

$$p_m = \lambda\,(m - 0.44) / (M + 1.12)$$

and the corresponding "empirical return period":

$$T_m = 1/p_m$$

b) Plot each x_i at its corresponding plotting position on exponential probability paper. Points should be relatively large - of the order of 1/80 of the minimum plot dimension. The legend should note "observed peaks over threshold". An example of exponential probabilty paper is provided in Appendix B6. If exponential probability paper is not available the data can be plotted on linear graph paper using the linear reduced variate, Rv, for the exponential distribution, defined as:

$$Rv = -\ln (p_m)$$

c) On the same exponential probability paper, plot the quantile estimates for a range of return periods as construction points . Join these construction points with a bold line labelled "fit line - exponential distribution".

4.3 Post Analysis Checking

In many cases, the existence of an apparent outlier may be suspected after the extreme value analysis is completed. An outlier may be suspected because one or more data points depart significantly from the trend of either the data or the fitted distribution, plotted as described above. An outlier may exist for one of several reasons:

- it may deviate from other observations because it was generated by some other process,
- it may be a rare event from the same population distribution,
- it may be an observational error which was not detected in pre-analysis examination,
- the statistical model may be wrong.

The extent to which an outlier may represent a problem is a function of the type of distribution, the sample size and the presence of corroborative regional information. It should be noted that:

- three parameter distributions are more sensitive to outliers than two parameter distributions,
- the influence of an outlier on the moments of a distribution are larger for a small sample than for a large sample,
- if the date of the outlier is known, raw or other data records can be reviewed for corroborative evidence which might support or explain the presence of the outlier.

The following procedures are recommended depending on whether the suspected outlier is a high outlier or a low outlier.

4.3.1 High Outliers

If the apparent outlier is a high outlier, that is, it has a low probability of occurrence, the following procedures should be undertaken:

- Determine if the data point is correct by checking the source or basis of the data for errors associated with either measurement or recording or archiving methodology, especially where the data was derived using numerical models.
- Verify the physical processes which are the underlying cause of the observation, and determine if it is consistent with the other observations.

If the outlier is the result of a non-homogeneous process it should be removed and consideration should be given to quantifying and combining the probabilities of two processes. If the apparent outlier is a homogeneous observation, the calculations should be rechecked. If they are correct, then:

- the observation may be a rare event, or,
- the model may be wrong, and consideration should be given to using another extreme value model.

If the model is not wrong or there is no other suitable extreme value model, then the best that one can do is to repeat the analysis without the outlier and define an envelope of likely return periods for the observation.

In the special case of Forensic analysis, discussed further in Chapter 5.3, a more detailed understanding of the outlier and its statistical significance is required.

4.3.2 Low Outliers

The occurrence of low outliers is an often overlooked phenomenon, but the user should review any analysis for the presence of such outliers. In particular low outliers can significantly distort sample skewness and the user will be inclined to use distributions which are consistent with low skewness (i.e. the Normal distribution or a three parameter distribution such as the GEV, which can have an upper limit). The danger of a low outlier is that it may create an artificial case where the estimated upper limit is less than the largest observation.

If a low outlier is suspected, the following procedures should be undertaken:

- Contact the original data collection agency to determine that the data points are correct.
- Verify the physical processes which are the underlying cause of the observation and determine if it is consistent with the other observations.
- If the outlier is the result of a non-homogeneous process it should be removed and consideration should be given to how the probabilities of two processes should be quantified and combined.
- If the apparent outlier is consistent with the remainder of the sample, the calculations should be rechecked and if they are correct, the model may be wrong. Consideration should then be given to using another extreme value model. If a three parameter distribution was used for the conditional probability distribution, the estimated upper band should be checked to ensure that it is consistent with the observed data.

4.4 Uncertainty of Extremes Estimates

The uncertainty due to sample size alone, of extremes estimates, can be approximated using the sampling standard error, $se(x_T)$. The sampling standard error can be combined with the normal distribution to estimate confidence limits about any estimate of a quantile, x_T.

When the threshold is fixed, and parameters are estimated by moments for the Poisson/exponential POT model, the sampling standard error $se(x_T)$ can be computed from:

$$se\ (x_T) = [1 + (\ln\lambda + y_T)^2]^{1/2}\ \beta\ /\ (\ln\lambda)$$

5.0 ADDITIONAL CONSIDERATIONS

During the development of these guidelines, several important topics which were beyond the scope of earlier phases of this study emerged, which on examination were not generally discussed in the available literature. These are discussed below. Guidance in the form of references are provided to assist the user, but it should be recognized that further development may be required before a general methodology can be recommended

5.1 Choice of Underlying Variable

Prior to undertaking an extreme value analysis, it is worthwhile developing an definition of the type of variable which is required for the purpose of the extreme value analysis. For instance, it is necessary to choose between variables which define a process, such as millimeters of rainfall, windspeed or wave height, and variables which define the response of a system or structure such as water reservoir capacity, stress in a structural element or damage to a shoreline or coastal structure.

In the first case, it is usually sufficient to undertake an extreme value analysis using readily available data from a data collection agency. In the second case, it may be necessary to account for the effect of several variables and either a suitable statistical model must be used, such as those referred to in Section 5.5, or a suitable response index or model must be developed which would characterize the system and its response to the processes of interest, (i.e. a wave hindcast model).

Selection of the variable for extreme value analysis should be undertaken with the assistance of appropriate expertise. Some guidelines in selecting appropriate variables are provided below:

- A primary reason for undertaking an extreme value analysis is to provide information to be used to develop a system or structure with a predictable and desirable response as it reaches the limits of its intended purpose. The selected variable should be consistent with the objective of quantifying this response.

- The return period of an event defines the median waiting interval between successive occurrences of the same event. This can be interpreted as the expected waiting time between similar responses of a particular system or structure. If the response leads to or includes failure, the expected waiting time will define the probability of failure. In many situations, failure is the primary concern underlying the need for an extreme value analysis and therefore the analysis should be undertaken using variables which quantify the system response of concern.

5.2 Data Uncertainty

Uncertainty exists in any statistical procedure and in an extreme value analysis uncertainty may be present for one or more of several reasons:

- the true population of events is unknown
- the sample is small

- the process of interest may be changing in a way which can not be detected from the small sample
- the individual data in a sample are only accurate within finite limits.

The uncertainty resulting from the lack of knowledge of the true population can not be quantified. The uncertainty can only be minimized by selecting appropriate extreme value models.

The uncertainty resulting from small samples can be quantified using the procedures defined in sections 3.4 and 4.4 or this report for the AMAX and POT models, respectively.

The uncertainty due to misinterpr\ ation of stationarity can not be quantified, but can be minimized with a through understanding of the processes of interest.

The statistical uncertainty due to random error in the data used for an extreme value analysis can not generally be quantified. If the error is randomly distributed, the sample variance (s_x) will be greater than the true sample variance, and the quantile estimates will be over-estimated. Earle and Baer (1982) quantified this effect for a lognormal distribution, but specific guidelines which define this effect for the extreme value models described in this report are not available.

5.3 Forensic Extreme Value Analysis

The term forensic extreme value analysis is used here to refer to the case in which the objective of the analysis is to estimate the return period of an event of specified magnitude. This event may be an actual observed event which was particularly severe and may have led to property damage or loss of life. It may also be an event that has been simulated using a deterministic model.

This type of analysis is different from typical extreme value analysis in which the objective is to estimate the magnitude of a particular event given a specific return period. Although many of the considerations are similar, the analysis can be complex and is beyond the scope of these guidelines.

In any event, the consequences of being wrong (or of appearing to be wrong) in the case of forensic analysis are such, that this type of problem should be referred to an expert.

5.4 Other POT models

The guidelines described in this report for the POT model are based on the use of the truncated exponential distribution for the conditional probability distribution of magnitudes (given that they occur), in a Poisson process. This form of the POT model is the most common version used and has complete and well defined procedures for the various stages of analysis, as described in this report.

Other versions of the POT model may prove to be equally valid, however, it was beyond the scope of this study to define, test and evaluate the many versions which exist. One useful version substitutes the truncated normal distribution for the truncated exponential distribution. In situations where the sample skewness is significantly smaller than 2, this model may prove useful. Details of the truncated normal distribution are provided in Kendall and Stuart (1977).

Caution should be exercised in substituting other probability distributions in the POT model because of the danger of overfitting the model by introducing additional parameters. This problem is similar to that faced in the AMAX model when a three parameter distribution is used with too small a sample size.

5.5 Multivariate Extreme Value Analysis

In some situations it may be desirable to define the extreme values of a specific process defined by more than one variable. The models described in this report are all univariate models and can be used in this type of application only if the joint effect of multiple variables is quantified through the use of a response index or model.

If the variables are mutually independent, quantiles can be estimated from the product of the individual exceedance probabilities from a univariate model. Similarly if they are totally dependent they can be defined by the exceedance probabilities of one of the variables. If the variables are only partially dependent, a multivariate extreme value model is required. Multivariate models for extreme value analysis have been developed recently and are described in Tiaga de Oliveria (1984).

Many of the multivariate models are relatively new and further development is required for practical applications. A multivariate extreme value analysis should be undertaken with the assistance of appropriate expertise.

REFERENCES

Earle, M.D. and Baer, L. "Effects of Uncertainties on Extreme Wave Heights", Journal of the Waterway, Port, Coastal and Ocean Division, ASCE, Nov 1982.

Hogg, W.D. and Carr, D.A., "Rainfall Frequency Atlas for Canada", Canadian Government Publishing Centre, Ottawa, 1985.

Kendall, M. and Stuart, A., "The Advanced Theory of Statistics", 3 Volumes, 4th Edition, McMillan Publishing Co., Inc., New York, 1977

Natural Environment Research Council, "Flood Studies Report", NERC, London, 1975.

Tiago de Olivera, J., Ed. "Proceedings of the NATO Advanced Study Institute on Statistical Extremes and Applications" Vimeira, Portugal, Riedel Publishing Co., Holland, 1984

W.F. Baird & Associates Coastal Engineers Ltd., and HYDROTEK Water Resources Consultants Inc., "Review and Assessment of Procedures for Extreme Value Analysis for Selected Geophysical Data, Phase I", Canadian Climate Centre Report 86-12, AES, Downsview, 1986.

W.F. Baird & Associates Coastal Engineers Ltd., and HYDROTEK Water Resources Consultants Inc., "Review and Assessment of Procedures for Extreme Value Analysis for Selected Geophysical Data, Phase II", Canadian Climate Centre Report 87-5, AES, Downsview, 1987.

APPENDIX A

EXAMPLE OF CASE A/AMAX

A.1 Data Preparation

An annual maximum series of hourly wind speed at a constant elevation has been provided by a data collection agency. The period of record is 56 years (1929-1984) and there are no missing data. A tabular listing of the data is provided in Table A.1.

Table A.1

Listing of Data

CHRONOLOGICAL		ORDERED			
Year	Windspeed (kph)	Windspeed (kph)	Rank	Plotting Position	Return Period (years)
1929	119.0	130.4	1	.011	93.667
1930	109.0	125.6	2	.028	35.125
1931	103.0	119.2	3	.046	21.615
1932	82.0	119.0	4	.064	15.611
1933	85.0	117.5	5	.082	12.217
1934	91.8	115.9	6	.100	10.036
1935	101.4	111.0	7	.117	8.515
1936	86.9	109.0	8	.135	7.395
1937	91.8	106.3	9	.153	6.535
1938	119.2	106.3	10	.171	5.854
1939	80.5	104.7	11	.189	5.302
1940	104.7	104.7	12	.206	4.845
1941	95.0	103.0	13.	.224	4.460
1942	104.7	103.0	14	.242	4.132
1943	93.4	101.4	15	.260	3.849
1944	96.6	99.8	16	.278	3.603
1945	125.6	96.6	17	.295	3.386
1946	106.3	96.6	18	.313	3.193
1947	86.9	96.6	19	.331	3.022
1948	90.2	96.6	20	.349	2.867
1949	96.6	96.6	21	.367	2.728
1950	106.3	96.0	22	.384	2.602
1951	93.4	95.0	23	.402	2.487
1952	88.6	93.4	24	.420	2.381
1953	77.3	93.4	25	.438	2.285
1954	77.3	93.4	26	.456	2.195
1955	91.8	93.4	27	.473	2.113
1956	77.3	93.0	28	.491	2.036
1957	93.4	93.0	29	.509	1.965
1958	86.9	91.8	30	.527	1.899
1959	83.7	91.8	31	.544	1.837
1960	77.3	91.8	32	.562	1.778

1961	90.2	90.2	33	.580	1.724
1962	78.9	90.2	34	.598	1.673
1963	96.6	89.0	35	.616	1.624
1964	96.6	88.6	36	.633	1.579
1965	117.5	88.6	37	.651	1.536
1966	103.0	86.9	38	.669	1.495
1967	93.4	86.9	39	.687	1.456
1968	115.9	86.9	40	.705	1.419
1969	83.7	85.3	41	.722	1.384
1970	96.6	85.0	42	.740	1.351
1971	82.1	85.0	43	.758	1.319
1972	85.3	85.0	44	.776	1.289
1973	82.1	83.7	45	.794	1.260
1974	99.8	83.7	46	.811	1.232
1975	88.6	32.1	47	.829	1.206
1976	130.4	82.1	48	.847	1.181
1977	85.0	82.0	49	.865	1.156
1978	78.0	80.5	50	.883	1.133
1979	93.0	78.9	51	.900	1.111
1980	85.0	78.0	52	.918	1.089
1981	96.0	77.3	53	.936	1.068
1982	111.0	77.3	54	.954	1.049
1983	93.0	77.3	55	.972	1.029
1984	89.0	77.3	56	.989	1.011

A.2 Pre-Analysis Examinations

The time series of the annual maxima is shown in Figure 3.2. Neither jump nor trend is evident so the tests outlined in Appendix C are not required. There also does not appear to be any dependence or clustering of the data, but as a check, the autocorrelation function (Appendix C4) is computed.

The values calculated for r_k, for k = 1 to 5 years, using the data in Table A.1, are:

0.10, 0.10, -0.03, -0.06, and -0.01.

The standard error of r_k is approximately 0.13. The values of r_k are within 2 standard errors of zero, which is the expected value of the autocorrelation function for an independent series.

A.3 Sample Moments

The sample moments of the data in Table A.1 are:

$\bar{x} = 94.5$ kph

$s_x = 12.8$ kph

$g_x = 0.9$

A.4 Graphical Presentation of the Observed Annual Maxima

The data are plotted on Figure A.1, using the plotting positions tabulated in Table A.1. For the purpose of economy, the fitted distribution selected in Appendix A.5 is also shown at this stage of the analysis.

A.5 Selection of Distribution

The sample skewness coefficent, $g_x = 0.9$ for N = 56. From Figure 3.3 this is within 1/2 a standard error of the population skewness coefficent for the EV1 distribution. Therefore assume an EV1 distribution.

A.6 Quantile Estimates for the EV1 Distribution

Based on section 3.2.5, the quantiles for the EV1 distribution are estimated using the following formula:

$$x_T = 94.5 + 12.8\ K_T$$

Values for the quantiles are provided in Table A.2.

A.7 Graphical Presentation of Quantile Estimates

The fitted distribution is shown on Figure A.1

A.8 Sampling Standard Error - EV1 Distribution

The uncertainty due to sampling of the quantiles summarized above can be estimated using the following expression described in section 3.4.1:

$$se\ (s_T) = (12.8 / \sqrt{56})\ \delta_T = 1.71\ \delta_T$$

Values of the standard error are provided in Table A.2.

Table A.2
Summmary of the Results of the Annual Maximum Model

Return Period (years)	Magnitude (kph)	Standard Error (kph)
1.05	77.7	2.0
1.25	84.1	1.5
2.0	92.4	1.6
5.0	104.0	2.6
10.0	111.0	3.6
20.0	118.0	4.5
50.0	128.0	5.8
100.0	135.0	6.7
200.0	141.0	7.7
500.0	151.0	8.9

A.9 Post-Analysis Examination

Reference to Figure A.1 shows that the EV1 fit line closely approximates the observed annual maxima and that there are no outliers.

APPENDIX B

EXAMPLE OF CASE B/POT

B.1 Data Preparation

A parent POT series corresponding to all hourly wind speeds over a threshold of 50 km/h has been provided by a data collection agency. The length of record is 10 years (1975 - 1984). The series was screened to yield an "independent" parent series with a minimum inter-event time of 24 hours. The total number of remaining events is 331. A histogram for this series is plotted in Figure B.1. The mode of the parent population is 25 km/h.

A second threshold of 68 km/h was selected which yielded a sample size (M) of 102. These data are summarized in Table B.1. Note that a larger threshold (and correspondingly smaller M) could have been selected. In this case, a threshold value was selected arbitrarily such that it produced on average approximately 10 storms per year.

Table B.1

Observed Data Over Threshold in Chronological and Ranked Form

CHRONOLOGICAL		ORDERED			
Year	Windspeed (km/h)	Windspeed (km/h)	Rank	Probability	Return Period (years)
75	71.7	130.50	1	.01	17.88
75	69.30	109.80	2	.02	6.42
75	90.40	101.30	3	.03	3.91
75	74.10	99.70	4	.03	2.81
75	71.70	98.10	5	.04	2.20
75	72.90	95.00	6	.05	1.80
75	75.40	93.40	7	.06	1.53
76	70.50	91.90	8	.07	1.32
76	77.90	91.90	9	.08	1.17
76	99.70	90.40	10	.09	1.05
76	74.10	89.00	11	.10	.95
76	70.50	89.00	12	.11	.87
76	68.10	89.00	13	.12	.80
76	70.50	87.50	14	.13	.74
76	130.50	86.10	15	.14	.69
76	89.00	86.10	16	.15	.64
76	68.10	86.10	17	.16	.60
76	76.70	86.10	18	.17	.57
76	76.70	86.10	19	.18	.54
76	70.50	86.10	20	.19	.51
76	70.50	84.70	21	.20	.49
76	74.10	84.70	22	.21	.46

76	74.10	83.30	23	.22	.44
77	80.60	83.30	24	.23	.42
77	69.30	83.30	25	.24	.41
77	68.10	83.30	26	.25	.39
77	74.10	83.30	27	.26	.38
77	86.10	81.90	28	.27	.36
77	68.10	81.90	29	.28	.35
77	83.30	81.90	30	.29	.34
77	75.40	81.90	31	.30	.33
77	86.10	81.90	32	.31	.32
77	77.90	80.60	33	.32	.31
77	70.50	80.60	34	.33	.30
77	79.30	79.30	35	.34	.29
77	74.10	77.90	36	.35	.28
78	70.50	77.90	37	.36	.27
78	71.70	77.90	38	.37	.27
78	77.90	77.90	39	.38	.26
78	77.90	77.90	40	.39	.25
78	68.10	76.70	41	.40	.25
78	76.70	76.70	42	.41	.24
78	74.10	76.70	43	.42	.24
78	69.30	76.70	44	.43	.23
78	69.30	76.70	45	.44	.22
79	81.90	76.70	46	.45	.22
79	76.70	75.40	47	.46	.22
79	68.10	75.40	48	.47	.21
79	68.10	75.40	49	.48	.21
79	74.10	75.40	50	.49	.20
79	91.90	74.10	51	.50	.20
80	81.90	74.10	52	.50	.19
80	75.40	74.10	53	.51	.19
80	83.30	74.10	54	.52	.19
80	69.30	74.10	55	.53	.18
80	83.30	74.10	56	.54	.18
80	69.30	74.10	57	.55	.18
80	71.70	74.10	58	.56	.17
80	86.10	74.10	59	.57	.17
80	80.60	74.10	60	.58	.17
80	69.30	74.10	61	.59	.17
80	69.30	74.10	62	.60	.16
81	87.50	74.10	63	.61	.16
81	70.50	72.90	64	.62	.16
81	89.00	72.90	65	.63	.16
81	70.50	72.90	66	.64	.15
81	74.10	71.70	67	.65	.15
81	76.70	71.70	68	.66	.15
81	71.70	71.70	69	.67	.15
81	83.30	71.70	70	.68	.14
81	95.00	71.70	71	.69	.14
81	68.10	71.70	72	.70	.14
81	71.70	71.70	73	.71	.14
81	74.10	71.70	74	.72	.14
82	74.10	70.50	75	.73	.13
82	93.40	70.50	76	.74	.13
82	101.30	70.50	77	.75	.13
82	81.90	70.50	78	.76	.13
82	76.70	70.50	79	.77	.13
82	98.10	70.50	80	.78	.13

82	71.70	70.50	81	.79	.12
82	109.80	70.50	82	.80	.12
82	81.90	70.50	83	.81	.12
82	83.30	69.30	84	.82	.12
82	72.90	69.30	85	.83	.12
83	86.10	69.30	86	.84	.12
83	72.90	69.30	87	.85	.12
83	91.90	69.30	88	.86	.11
83	77.90	69.30	89	.87	.11
83	69.30	69.30	90	.88	.11
83	74.10	69.30	91	.89	.11
83	84.70	69.30	92	.90	.11
83	86.10	69.30	93	.91	.11
84	69.30	68.10	94	.92	.11
84	71.70	68.10	95	.93	.11
84	86.10	68.10	96	.94	.10
84	74.10	68.10	97	.95	.10
84	68.10	68.10	98	.96	.10
84	89.00	68.10	99	.97	.10
84	75.40	68.10	100	.97	.10
84	81.90	68.10	101	.98	.10
84	84.70	68.10	102	.99	.10

B.2 Pre-Analysis Examination

The time series of the peak over threshold events is shown in Figure 4.1 and it does not indicate any jump or trend so the tests outlined in Table 4.1 are not required. There also does not appear to be any dependence in the data with the exception of a tendency to cluster in the late fall and winter months of each year. Because this is wind data and the winds are strongest in the fall and winter months, this clustering does not indicate any statistical correlation.

B.3 Calculation of Sample Moments

The sample moments of the data in Table B.1 are:

$\bar{x} = 78$ kph
$s_x = 10.1$ kph
$g_x = 2.1$

B.4 Definition of the POT Parameters

The number of years of data, N = 10.

The number of selected events, M for a threshold of 68 kph = 102

Therefore the average number of events per year, $\lambda = 102 / 10 = 10.2$

The scale parameter $\beta = 78 - 68 = 10$

For these parameters , the ratio of the mean to the variance for the number of events/year is 1.13. This supports the use of the Poisson distribution. The sample skewness of 2.1 is close to 2 which supports the use of the expontial distribution. Therefore, the POT model is assumed to be valid and quantiles can be estimated.

For the purpose of illustrating the effect of the threshold on the parameters of the POT model, the threshold is increased incrementally with the results indicated in Table B.2.

Table B.2

Evaluation of POT Model

x_0 (km/h)	$\bar{x}$ (km/h)	g_x	λ	β	λ/var λ
68	78	2.1	10.2	10	1.1
72	82.5	2.3	6.6	10.5	1.3
80	89.1	2.7	3.4	9.1	0.8
85	93.7	2.4	2.0	8.7	1.3

The results indicate that the POT model is valid for all values of the threshold and the arbitrary choice of any threshold would have produced acceptable results within the limitations of statistical uncertainty.

B.5 Quantile Estimates

Based on section 4.2.3, the quantiles for the POT distribution are estimated using the following fomula:

$$x_T = x_0 + \beta \ln\lambda = \beta y_T$$

For example:

$$x_{100} = 68 + 10 \ln (10.2) + 10 (4.60) = 137 \text{ km/h}$$

A range of quantile estimates for various return periods is provided in Table B.3. for the threshold of 68 km/h.

Table B.3

Pot Quantile Estimates and Standard Errors

Return Period	Wind Speed	Standard Error
T	x_T	$se(x_T)$
(years)	(km/h)	(km/h)
1.005	75	1.2
2	95	2.8
5	106	3.9
10	114	4.6
20	121	5.3
50	130	6.2
100	137	6.9

B.6 Graphical Presentation

The fitted distribution in shown in .Figure B.2.

B.7 Sampling Standard Error - POT Approach

The uncertainty due to sampling of the quantiles summarized above can be estimated using the following expression described in section 4.4:

$$se\,(x_T) = \frac{B}{\sqrt{N}}\ [\,1 + (\,1n\,\lambda + y_T)^2\,]^{1/2}$$

$$= \frac{10}{\sqrt{10}}\ [1 + (1n\ 10.2 + y_T)^2\,]^{1/2}$$

The standard error of estimate is provided in Table B.3 for a range of return periods

APPENDIX C

STATISTICAL TESTS

C.1 Spearman Rank Order Correlation Coefficient Test for Trend

The significance of any trend can be tested using the Spearman Rank Order Correlation Coefficient, r_s (Kendall and Stuart, 1977). The test statistic r_s is a function of the difference in rank, d_i, between adjacent items in two series, $X_1 \ldots X_n$, and $Y_1 \ldots Y_n$. Here d_i denotes the difference in rank between a series arranged in order of decreasing magnitude and a series arranged in chronological order. If two or more observations in a series are equal, the mean rank is assigned to each tied observation. If the proportion of ties is not large, their effect on r_s is negligible and, the test statistic can be computed as

$$r_s = 1 - (6 \Sigma d_i^2) / (N^3 - N)$$

For sample size N of 10 or larger, the statistic, r_s, is distributed by Student's t distribution with N – 2 degrees of freedom. Therefore:

$$t = r_s [(n - 2) / (1 - r_s^2)]^{1/2}$$

A two-tailed test is used and the null hypothesis is that there is no trend.

If a significant number of tied ranks exist, a correction factor must be incorporated in the computation of r_s.

The correction factor for tied ranks is given by:

$$S = \Sigma(t_m^3 - t_m)/12$$

where t_m is the number of observations tied at a given rank. If R_x refers to the ranks of the series $X_1, X_2 \ldots X_N$ and R_y to the time series represented by $Y_1, Y_2 \ldots Y_N$, then the test statistic is given by:

$$r_s = (\Sigma R_x^2 + \Sigma R_y^2 - \Sigma d_i^2) / [2 \sqrt{(\Sigma R_x^2 \, \Sigma R_y^2)}]$$

where:

$$\Sigma R_x^2 = [(N - 1)^3 - (N - 1)] / 12 - S_x,$$

$$\Sigma R_y^2 = [(N - 1)^3 - (N - 1)] / 12 - S_y,$$

S_x and S_y are the correction factors for ties for R_x and R_y respectively.

C.2 Wald-Wolfowitz Test for Jump

The Wald-Wolfowitz split sample test, (Kendall and Stuart, 1977) can be used to determine whether two samples have significantly different means, variances, skewness or kurtosis. The test statistic is the number of runs, R_{ww} , for either subseries, in the rank series for the combined and ranked sample series . For sample sizes $n_1 > 4$ and $n_2 > 20$, the sampling distribution tends to normality with z a standard normal variate.

$$z = \frac{R_{ww} - [(2n_1n_2) / (n_1 n_2) + 1] - 0.5}{[2n_1 n_2 (2n_1 n_2 - n_1)] / [(n_1 = n_2)^2 (n_1 + n_2 - 1)]^{1/2}}$$

C.3 Mann-Whitney Test for Jump

The Mann-Whitney test statistic for jump is the same as that for homogeneity except that the subsamples are collected before and after the suspected jump (Appendix C.6). The Mann-Whitney statistic U, is used to determine whether the subsamples have significantly different means.

C.4 Autocorrelation Coefficient Test for Independence

Even if a variable is random and stationary, it may not be independent. For streamflow, large natural storage such as lakes or groundwater reservoirs may cause large flows to be followed by more high flows, while in meteorology, large scale features of the earth's atmospheric circulation may cause a succession of severe storms.

Such persistence can be reflected by the autocorrelation function Q_k which, for a sample of size N is estimated by its sample value, r_k.

$$r_k = N \sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x}) / (N-k) \qquad \sum_{i=1}^{N} (x_i - \bar{x})^2$$

For large samples, the function is approximately normally distributed with variance 1/N and hence a confidence band of 0 plus or minus $2/\sqrt{N}$ would include all estimates within two standard errors of the expected value of zero for an uncorrelated time series.

C.5 Spearman Rank Order Serial Correlation Coefficient Test for Independence

The Spearman rank order serial correlation coefficient can also be used to test for independence.

The test statistic r_s is the same for Spearman test for trend except that R_x refers to the ranks of the series $X_1, X_2, ..., X_{N-1}$, and R_y to the ranks of the series $X_2, X_3, ..., X_N$. If two or more observations have equal values, the same correction factor must be applied.

If the proportion of ties is not large, their effect on r_s is negligible, that is, $S << [N-1)^3 - (N-1)]/12$, and r_s can be computed using.

$$r_s = 1 - (6 \Sigma d_i^2) / [(N-1)]^3 - (N-1)]$$

The quantity:

$$t = r_s[(N - 3) / (1 - r_s^2)]^{1/2}$$

is distributed as Student's t with N–3 degrees of freedom when $N > 11$.

Recent developments indicate that in many cases it is not essential that data series be independent and that the classical extreme value theory also applies to stationary dependent series. Some of the results for the dependent case may also be extended to non-stationary situations.

C.6 Mann-Whitney Test for Homogeneity

The Mann and Whitney test, (Kendall and Stuart, 1977) is used to determine whether two samples have significantly different means. The Mann-Whitney test statistic, U, is a function of the sum of the ranks and is defined by the smaller of:

$$U_1 = n_1 n_2 + n_1 + n_1 (n_1 + 1) / 2 - R_1$$

or $$U_2 = n_1 n_2 - U_1$$

Here R_1 denotes the sum of the ranks in subsample 1.

For $n_1 > 4$ and $n_2 > 20$, the sampling distribution of U tends rapidly to normality with

$$z = \frac{U - n_1 (n_2 / 2)}{\left\{\left[\frac{n_1 n_2}{N(N-1)}\right]\left[\frac{N^3 - N}{12} - S\right]\right\}^{1/2}}$$

where:

$$S = \Sigma(t_m^3 - t_m)/12$$

where t_m is the number of observations tied at a given rank.

The samples can be said to have a significant difference of means if:

$$z < -1.645 \text{ for significance level } \alpha = .05$$

or

$$s < -2.326 \text{ for } \alpha = .01.$$

For n_2 up to 20, the critical values of U for significance levels A = .05 and A = .01 are tabulated in Tables C.1 and C.2. If the computed value of U is less than or equal to the tabulated values, the samples are said to have significant difference of means at the level indicated.

C.7 Wald-Wolfowitz Test for Homogeniety

The Wald-Wolfowitz test statistic for homogeneity is the same as that used for jump except that the subsamples are sorted to conform to the suspected processes. The Wald-Wolfowitz statistic Rww is used to determine when the subsamples have significantly different means.